INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Pancerasa M, Sangiorgio M, Ambrosini R, Saino N, Winkler DW, Casagrandi R. 2019 Reconstruction of long-distance bird migration routes using advanced machine learning techniques on geolocator data. *J. R. Soc. Interface* **16**: 20190031.
http://dx.doi.org/10.1098/rsif.2019.0031

Received: 16 January 2019
Accepted: 29 May 2019

Subject Category:
Life Sciences – Engineering interface

Subject Areas:
bioinformatics

Keywords:
light-level tag, movement ecology, migratory species, path estimation, random forest, deep neural network

Authors for correspondence:
Mattia Pancerasa
e-mail: mattia.pancerasa@polimi.it
Matteo Sangiorgio
e-mail: matteo.sangiorgio@polimi.it
Renato Casagrandi
e-mail: renato.casagrandi@polimi.it

†These authors contributed equally to these study.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4531727.

THE ROYAL SOCIETY PUBLISHING

Reconstruction of long-distance bird migration routes using advanced machine learning techniques on geolocator data

Mattia Pancerasa[1,†], Matteo Sangiorgio[1,†], Roberto Ambrosini[2], Nicola Saino[2], David W. Winkler[3] and Renato Casagrandi[1]

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Giuseppe Ponzio, 34, Milano 20133, Italy
[2]Department of Environmental Science and Policy, Università degli Studi di Milano, Via Celoria 26, Milano 20133, Italy
[3]Department of Ecology and Evolutionary Biology, Cornell University, Corson Hall, Ithaca, NY 14853, USA

MP, 0000-0003-2654-7074; MS, 0000-0003-1624-6809; RA, 0000-0002-7148-1468; DWW, 0000-0003-4592-2546; RC, 0000-0001-5177-803X

Geolocators are a well-established technology to reconstruct migration routes of animals that are too small to carry satellite tags (e.g. passerine birds). These devices record environmental light-level data that enable the reconstruction of daily positions from the time of twilight. However, all current methods for analysing geolocator data require manual pre-processing of raw records to eliminate twilight events showing unnatural variation in light levels, a step that is time-consuming and must be accomplished by a trained expert. Here, we propose and implement advanced machine learning techniques to automate this procedure and we apply them to 108 migration tracks of barn swallows (*Hirundo rustica*). We show that routes reconstructed from the automated pre-processing are comparable to those obtained from manual selection accomplished by a human expert. This raises the possibility of fully automating light-level geolocator data analysis and possibly analysing the large amount of data already collected on several species.

1. Introduction

Recent decades have been characterized by environmental and climatic changes that occur on the global scale [1]. Ecological systems are responding to such changes with shifts in distribution and changes in the timing of ecological events [2]. Migratory animals are considered particularly sensitive to global changes because they should adjust their life cycle to changes that occur at different rates in areas that are separated by long distances [3,4]. It is, therefore, particularly important to understand if and how the movement pattern of long-distance migratory species is affected by climate and environmental change.

The tracking of migratory animals is a very active field of biological study [5–9]. Knowing the positions visited and the routes travelled by migratory organisms is in fact crucial information, for instance, to design management policies for species conservation [10,11] or for managing disease spread [12]. This research field is particularly challenging and very interdisciplinary, because it requires the integration of knowledge coming from biology and environmental science with that coming from IT engineering. The continuous development of new devices—which allow monitoring, recording and sometimes transmitting the positions of individuals over long time periods and large spatial extents—opens novel research perspectives that must be accompanied by similarly advanced ways of interpreting the newly available data through proper modelling and software.

Migratory birds are ideal organisms for enhancing the current research in the field, since an incredible variety of movement patterns is offered to investigation: every year, billions of individuals of several species make extensive

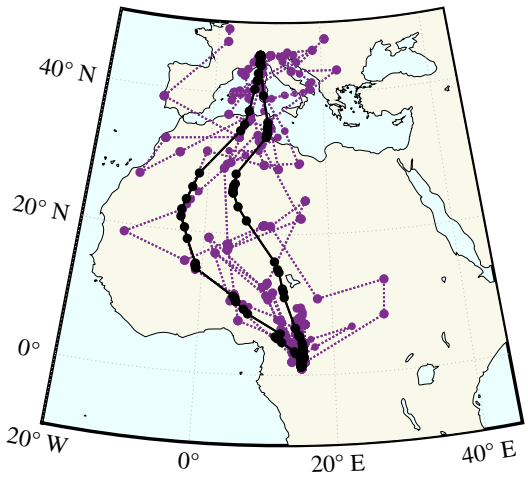

Figure 1. Comparison between a migration track estimated with FLightR using uninterpreted twilight data (dotted purple) and expert-classified twilight data (solid black). Filled circles represent the estimated bird position at each twilight event. (Online version in colour.)

journeys, covering astonishing distances over land or ocean and arriving at their destinations with impressive precision [13]. Some of these movements can be tracked with high-precision systems, such as GPS or other satellite receivers. The necessary equipment, however, is still too heavy to be carried by individuals of small species, for which every fraction of a gram of extra weight from tracking devices can make a huge difference in terms of success or failure of their migration event and their survival [14,15].

Thanks to their small size and minimal weight, light-level geolocators represent a valid, well-established technology to track movements of long-distance migratory animals [16–18]. These devices record the solar irradiance at regular time intervals, in the order of minutes. Starting from these data, it is possible to infer the geographical positions visited by an individual during its migration journeys using methods based on times of sunrise and sunset developed in the last decade [19,20]. However, a fundamental step required by all available methods is a sort of pre-processing of raw light-level data before treatment. This is because light measurements can be affected by shading effects due to different causes (e.g. cloudiness or bird resting in a dark cavity) that, if not filtered out, introduce errors into any method for estimating positions.

Recent programs for reconstructing migration routes from geolocator data (e.g. FLightR, [21]), rely on light-level values measured around twilight events (i.e. the 'template fit' method, developed by Ekstrom [22]) for estimating positions. Hence, each recorded twilight event showing unnatural variation of light levels around twilight (e.g. too abrupt changes of light levels in short periods of time or non-monotonic changes of light levels near twilights) needs to be manually removed by an expert performing a visual selection. If this operation is not performed correctly, the reconstruction of the routes is strongly influenced by noisy twilight events, which are responsible for producing highly biased estimates of the geographical positions (figure 1).

Since the track of one geolocator usually encompasses the movements of an individual for about 10 months, this manual selection requires the by-eye inspection of at least 600 twilight events, which is quite a cumbersome and time-consuming task. Some R packages have been developed to assist the expert during this work. The software TwGeos [23], for example, automatically identifies twilight events from raw light data and easily displays light variations occurring during all sunrises or sunsets (figure 2).

However, the selection of twilight events needs to be performed manually and it is still quite a slow and delicate operation, as on average an expert can classify the twilight events from no more than a handful of geolocators per day. The expert must discriminate noisy (shadowed) from natural twilight events by inspecting two different patterns: the variation of light intensity values after sunrise/before sunset as measured by the geolocators on the focal day i, and the smoothness of the day-to-day variation of sunrise and sunsets times around day i. From a statistical learning point of view, this problem can be interpreted as a binary classification (keep versus discard) of each twilight event to be performed using a set of numerical variables (predictors).

The specific aim of the present work is to develop a procedure that could automate twilight event classification by the implementation of machine learning (ML) algorithms. To build a substantial training dataset, we used more than 100 geolocator tracks of barn swallows during autumn and spring migration, for which we had previously manually classified about 40 000 twilight events. This large reference dataset allowed us to apply, in addition to a standard linear classifier (i.e. logistic regression, LR), also two ML classification algorithms (random forest, RF, and deep neural network, DNN). We selected the inputs to the classifiers from the geolocator light measurements and from the times of the previously identified twilight events. Finally, we tested the reliability of classifications by the different algorithms by estimating the migration routes travelled by four target individuals from twilights classified by each of the different algorithms and comparing the resulting routes to the one estimated from the expert-selected twilights.

2. Material and methods

2.1. Processing of geolocator data

We relied on data from 65 SOI-GDL2.10 and 43 SOI-GDL2.11 geolocators (Swiss Ornithological Inst., http://www.vogelwarte.ch/indirect-trackinggeolocator.html), 101 of which were used in the ornithological study of Liechti *et al*. [24] to analyse the migration of three populations of Swiss and Italian barn swallows (*Hirundo rustica*) between 2010 and 2012. SOI-GDL geolocators detect and store light intensity every 5 min by assigning an integer value on a scale between 0 (full dark) and 64 (maximum detectable luminosity). Each geolocator in the dataset registered an average of 381 (with a standard deviation of 144) twilight events, providing a total of 39 572 twilight events. We used TwGeos R package to pre-process raw data, identifying twilight events timing, manually inspecting every suggested twilight event and discarding those that showed light curves that were too different from the 'natural' ones (i.e. those that would be measured if the geolocator was placed in an open air location at the same geographical position, figure 2) and/or whose timing differed in an inconsistent way from that of the corresponding events occurring the days before or after the focal date.

Because the classification of certain twilights could be somewhat uncertain even for an expert (i.e. the light curve shows an intermediate pattern between the natural variation of accepted twilights and the luminosity increases in shadowed events), the same expert performing the first classification (labelled EXP1

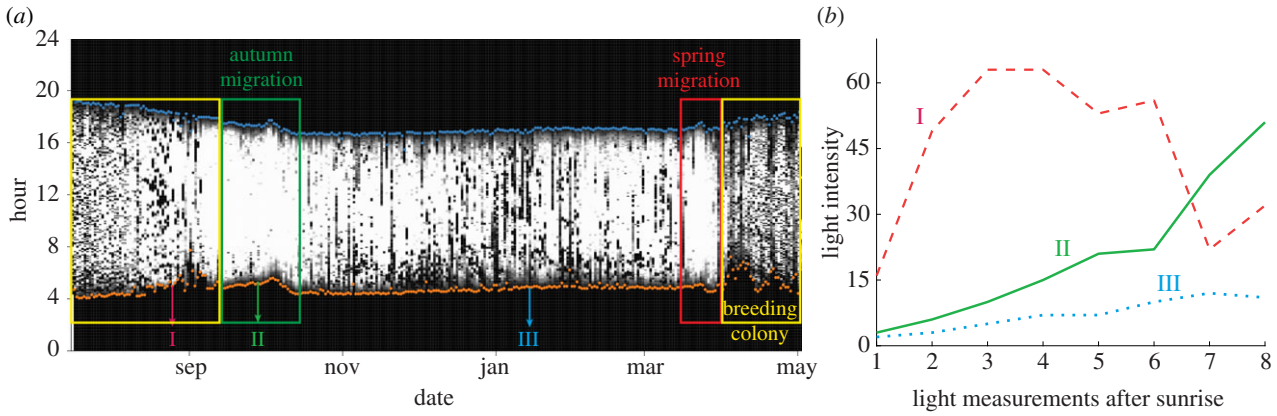

Figure 2. Light measurements of a light-level geolocator. (*a*) Geolocator raw data pre-processed by TwGeos R package, *x*-axis represents the day of the year, *y*-axis is the hour of the day. Values on the grey scale represent light intensity measured by the geolocator. Three probable phases of barn swallow annual phenological cycle are highlighted: autumn migration (green rectangle), spring migration (red rectangle) and the spring/summer stay at the breeding colony (yellow rectangle). (*b*) Light intensity patterns of three distinct twilight events (sunrises) taken as examples from (*a*): a natural variation of light intensity at sunrise (curve II, solid green) is contrasted with unnatural variations due to their unacceptably rapid increase in light (curve I, dashed magenta) or shadowed data (curve III, dotted cyan). (Online version in colour.)

above) re-classified after few weeks the twilight events registered by the geolocators of four target individuals cited in the Introduction (this dataset of the 'second classification by the same expert' will be named EXP2). The twilight events of these four 'example' geolocators were left out of the training dataset. We used the example set of twilight data to assess the repeatability of the expert classification and the ones performed by the ML algorithms by assessing the values of intraclass correlation coefficient (ICC) [25], usually defined, in the framework of random effects models, as the proportion of the total variance accounted for by differences among groups (aka classes, in statistical terminology). In general, for a model with only one random effect, the ICC is equal to

$$\mathrm{ICC} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_R^2},$$

where σ_G^2 is the between-groups (i.e. classification method) variance and σ_R^2 is the residual variance of the model. Repeatability can be interpreted as the expected within-group correlation among measurements [26]. In our case, the response variable is the binary twilight selection ('keep' = 0, 'discard' = 1), while the geolocator-ID and the twilight event ID are inserted as nested random effects (twilight-ID nested within geolocator-ID). We computed ICC values from a generalized linear mixed model assuming a binomial data distribution with logit link function, built with lme4 R package [27].

2.2. Feature selection

The expert classification of thousands of twilight events showed that the most telling evidence guiding the expert's decision was based on the shapes of light curves in close proximity to each twilight event. Therefore, we first selected as relevant features for our ML algorithms the eight light measurements of each geolocator either following each sunrise or preceding each sunset. For the large majority of twilight events, it was sufficient for the expert to inspect these eight values in order to discriminate between reliable (natural) and unreliable (shadowed) data. Regular monotonic increases (dawn) or decreases (dusk) of light intensity after dawn or before dusk are a necessary, but not sufficient, condition for recognizing natural twilights. Indeed, some dusks and dawns had to be more closely evaluated by the expert, who contrasted the twilight times with those recorded in nearby days (figure 2). As additional input variables for our classifier, we therefore added the timing of all twilight events occurring in a 9-day window centred on the focal event (i.e. day $i \pm 4$ days). To further improve the ability of the algorithm to capture the procedure used by the expert in the classification process, we added the following four additional input variables that qualify different properties of the twilight hours in surrounding days:

(1) the time difference between the twilight timing of the focal day i and the average twilight timings of top 50% of earliest sunrises/latest sunsets in a moving window of 9 days centred on the focal day (i.e. day $i \pm 4$ days). The aim of this variable was to identify the twilight events that occurred at very different times from neighbours (generally late sunrises and early sunsets, figure 2);
(2) same as (1), but with the average twilight timing computed over a mobile window of 19 days (i.e. day $i \pm 9$ days) to identify outliers on a wider time span;
(3) the standard deviation of night duration on a mobile window of 9 days (i.e. day $i \pm 4$ days). This variable was created to easily identify the twilight events of the periods in which a bird was at its breeding colony, because patterns of sunrise and sunset times were very variable in that period (see again figure 2) due to the shadowing caused by the buildings where the barn swallows rest during the night. The expert classification discarded many of these twilight events;
(4) the residual of a linear regression of twilight timings (either sunrises or sunsets) on date in a mobile window of 9 days (i.e. day $i \pm 4$ days). This variable was built to identify outliers during the migration periods, when the sunrise and sunset times change considerably, but regularly, from one day to another (see again figure 2).

2.3. Machine learning classifiers

Using ML terminology, our task is a supervised binary classification learning problem, where the set of twilights must be partitioned into two groups based on characteristics of the features of each element. The algorithms used to train classifiers require a dataset where each sample is qualified by its features and is already categorized. Starting from these data, supervised learning techniques can be used to train a model predicting which group each twilight belongs to.

First, we implemented a simple linear classifier, LR, which splits the high-dimensional feature space with a hyperplane and classifies each sample based on its position relative to a

linear decision boundary [28]. This simple classifier was used as benchmark for a fair comparison with the other more articulated models.

Next, we introduced two advanced nonlinear ML models which efficiently deal with problems where classes are not linearly separable: an RF [29] and a DNN [30].

An RF is an ensemble of classification trees: each tree is the result of a recursive partition of the feature space, here performed by a Boolean test on a single variable at each node. As a direct consequence, the feature space in each tree is separated by orthogonal hyperplanes, which results in a box-like decision [31]. Each classification tree is made on a random (both on instances and features) subset of the training dataset. The algorithm that builds the tree operates with a top-down procedure, choosing at each step the variable that performs the best split of the data using an evaluation function. In particular, we implemented the RF using the Gini impurity as the evaluation function, a standard method to evaluate partitioning in tree-based algorithms. Given a set of objects of different classes, Gini impurity is defined as the probability of obtaining two objects of different classes by a random sample on the set [32].

A DNN is an algorithm that repeatedly performs a nonlinear transformation of the feature space (each hidden layer performs one transformation) and, at the end, splits the transformed multi-dimensional space with a hyperplane, as LR does. The DNN has multiple layers, each of which is composed of one or more nodes (neurons). The neurons of the first layer correspond to the features. There are different ways to connect two subsequent layers: the most used way is *fully connected*, also called *dense*, where each node is computed as a nonlinear function of all nodes of the previous layer. Alternative structures to fully connected layers are convolutional layers. In this case, each node is computed as a nonlinear function of a subset of neighbouring nodes of the previous layer (i.e. using a mobile window to select the nodes that will connect to the neuron of the following layer). For the structure of our DNN, we used dense layers, and we also explored and tested convolutional layers, which process, with one-dimensional (1D) nonlinear filters, each of the features of the three time series of the focal twilight (i.e. the eight after sunrise/pre-sunset light measurements and the two series of the timing of the nine sunrises and sunsets around the specific twilight event), since they are known to be efficient in dealing with time series for other problems [33]. The convolutional filters extract multiple new features, whose definition is optimized during the training process. These new features, together with the four combined features selected by the expert, are given as input to the last fully connected layers of the DNN, which finally performs the classification (figure 3). As usual for convolutional networks, an additional nonlinearity is included in each layer with a *Max Pooling* process, which applies a piecewise maximum operator at the output of the convolutional filters. At the end of the whole structure, the *Softmax* activation function normalizes to 1 the values in the output layer [34].

Thanks to the large amount of data available, we decided to randomly split the entire dataset into three sets, for training (70%), cross-validation (15%) and test (15%), respectively. The cross-validation dataset was used to select the best hyperparametrization of each model (e.g. the maximum number of splits of the RF or the number of hidden layers of the neural network), while the test set was devoted to compare the performances of the resulting models of each class (LR, RF and the different architectures of the DNN).

We did not adopt more advanced model validation techniques, such as *k*-fold or leave-one-out cross-validation, as we did not note any overfitting problems during any phase of the work. We also did not use any precaution to train LR and RF (the former has a very low number of parameters compared to the task complexity, the latter is an algorithm that very unlikely overfits data if a sufficient number of trees is used; [35]). By contrast, we used an early stopping criterion and a regularization for the same procedure in DNN. We implemented and trained all ML algorithms using the *Python* packages SCIKIT-LEARN [36] and KERAS [37]. The pseudocode of the whole analysis is reported in electronic supplementary material, appendix S1.

2.4. Estimation and comparison of migration routes

We used FLightR R package to estimate the migration routes of the four example individuals. This software was designed to reconstruct migration routes of birds from geolocator light measurements. FLightR is based on a hidden Markov chain model, obtained by merging a physical observational model of the light variation (i.e. using astronomical equations to get a likelihood for each geographical position from repeated measurements of light during twilight events) with an uncorrelated random walk model of bird movement (for further details, see [38]). To get a posterior estimate of the state of the hidden Markov model (i.e. the geographical distribution of the position of the tagged bird), FLightR uses a particle filter [39], a Monte Carlo algorithm that performs particularly well with nonlinear hidden models and noisy measurements (i.e. the light values). At the end of its run, the particle filter provides an estimate of the central tendency and of the associated uncertainty of the route travelled by the bird.

For each method and for each individual, we generated a TAGS file by linking the geolocator measurements with the output of the different twilight classification methods and used this file as input in the FLightR analysis. The resulting paths (LR, RF, DNN and second classification by the expert: EXP2) were compared to those obtained from the first classification by the expert (EXP1), both visually and by calculating—on spring and autumn separately—the *one-way distance* (OWD, [40]) between migration paths. For two migration paths A and B, the OWD is calculated as follows: for each point of route A, the algorithm computes the distance between it and its 'corresponding position' on route B (defined as the position of B at the minimum great circle distance from the focal point in A); then it obtains $\mathrm{OWD}_{A\to B}$ as the sum of the previous distances divided by the length of route A. The final OWD measure is defined as $\frac{1}{2}(\mathrm{OWD}_{A\to B} + \mathrm{OWD}_{B\to A})$.

3. Results

3.1. Bias and variance trade-off

The confusion matrix for the three ML models used in this study (figure 4) revealed that the performances obtained in the three phases of the model calibration (training, validation and test) were nearly the same. We therefore concluded that we avoided overfitting. The RF and the DNN performed significantly better than the simpler LR classifier, but they performed similarly to each other.

The absence of overfitting, however, does not guarantee that the ML models would perform similarly to a human expert. Figure 5 compares the performances of the ML algorithms on the classification of the example individuals with the results that we obtained in the second manual classification (performed by the same human expert who classified data for training ML algorithms). It can be noted that the selectivity (true negative/ground truth negative) of the RF and the DNN is comparable with that of the expert, while the recall (true positive/ground truth positive) of the two algorithms is slightly lower than that of the expert.

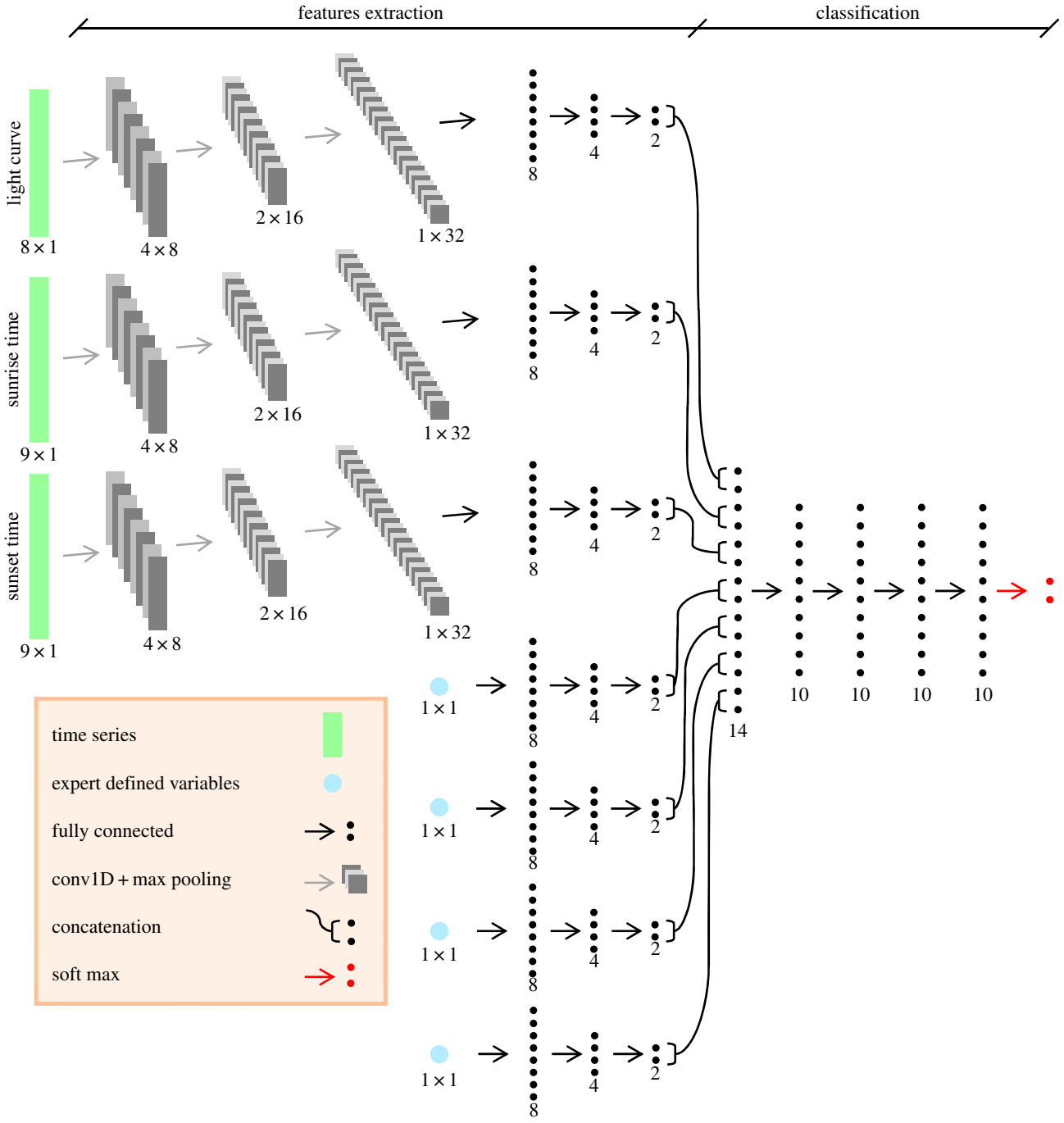

Figure 3. Schematic structure of the DNN used in our study. A first part of feature extraction is performed on the input variables. The three time series (green vectors: light curve at twilight event, sunrise time and sunset time of nearby days, respectively) are processed with 1D convolutional filters and then by a fully connected neural network (black dots: neurons). The four expert-defined features (four cyan single inputs) are not processed with convolutional filters and are directly submitted to the fully connected structure. (Online version in colour.)

Even in terms of overall accuracy, the LR (74.3) performed worse than the other two classifiers (RF: 87.6; DNN: 88.4), which in turn have a slightly lower score relative to the second classification by the expert (EXP2: 90.7). This last score means that the same expert doing the second classification (EXP2) provided a different classification than the one given in his first attempt (EXP1) in 9.3% of cases.

Table 1 shows the ICCs calculated from the results of the different classification methods. The ICC value from a mixed model that included the results of all classification methods is quite high ($ICC_{ALL\ versus\ ALL} = 0.82$), suggesting a generally good agreement between the expert classification and results of the ML models. The expert classification was the most consistent, as shown by the ICC between his two classifications ($ICC_{EXP1\ versus\ EXP2} = 0.87$). Among the other methods, RF and DNN had similar performances and outperformed LR. We can therefore state that both classification procedures operated by the expert and by ML algorithms are repeatable. The ICC at the geolocator level was never higher than 0.0114, implying that the classification of the twilight events did not depend on the features of an individual geolocator or of the barn swallow that carried it (i.e. it did not occur, for instance, that twilight events from one geolocator were consistently better than those of another). This means that the classification of twilight events was almost independent of the specific geolocator. This result was not strictly predictable as, in principle, individual barn swallows may differ in their behaviour during twilight events (e.g. they may consistently stay in more shaded

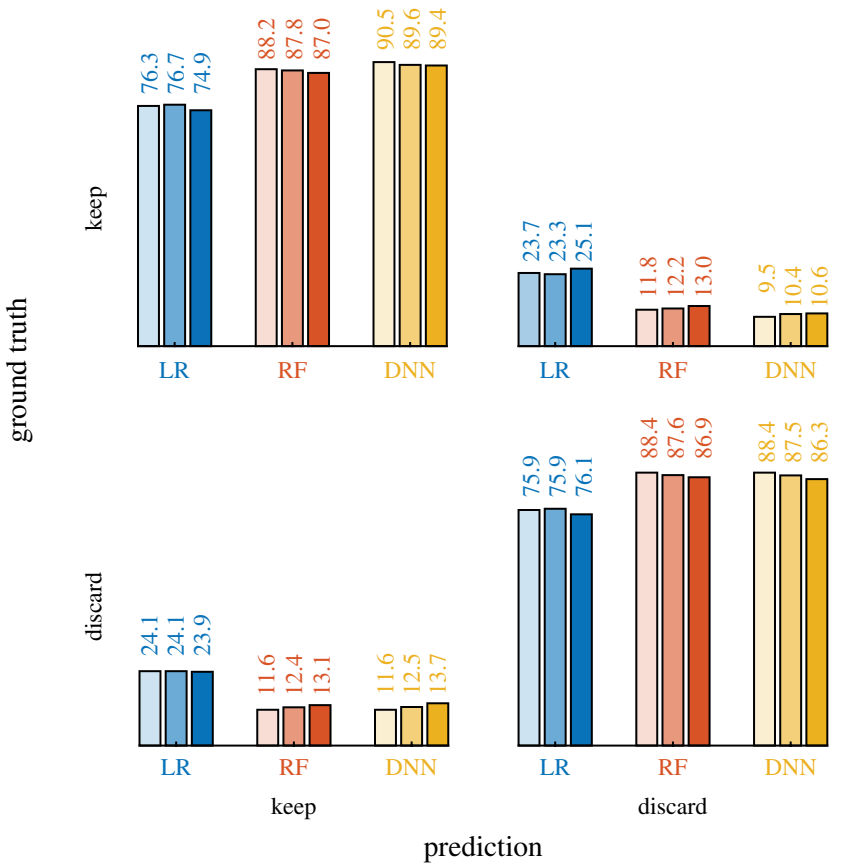

Figure 4. Confusion matrix of the three classifier approaches: LR (blue), RF (orange) and DNN (yellow). Performances are showed for training (light), validation (medium) and test (dark) sets. Ground truth is referred to the first expert classification (EXP1). (Online version in colour.)

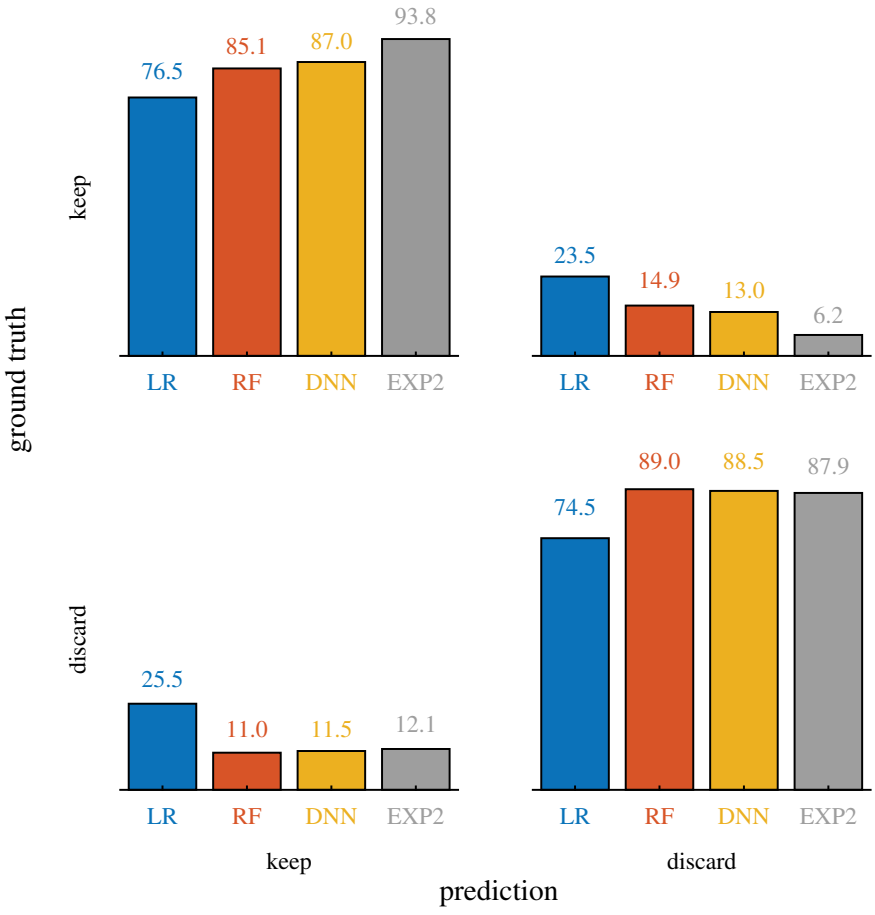

Figure 5. Comparison of classifications performed by our machine learning algorithms (LR, blue; RF, orange; DNN, yellow) and a human expert (EXP2, grey) on the twilight data of the target individuals. The ground truth comes from the first expert classification (EXP1). (Online version in colour.)

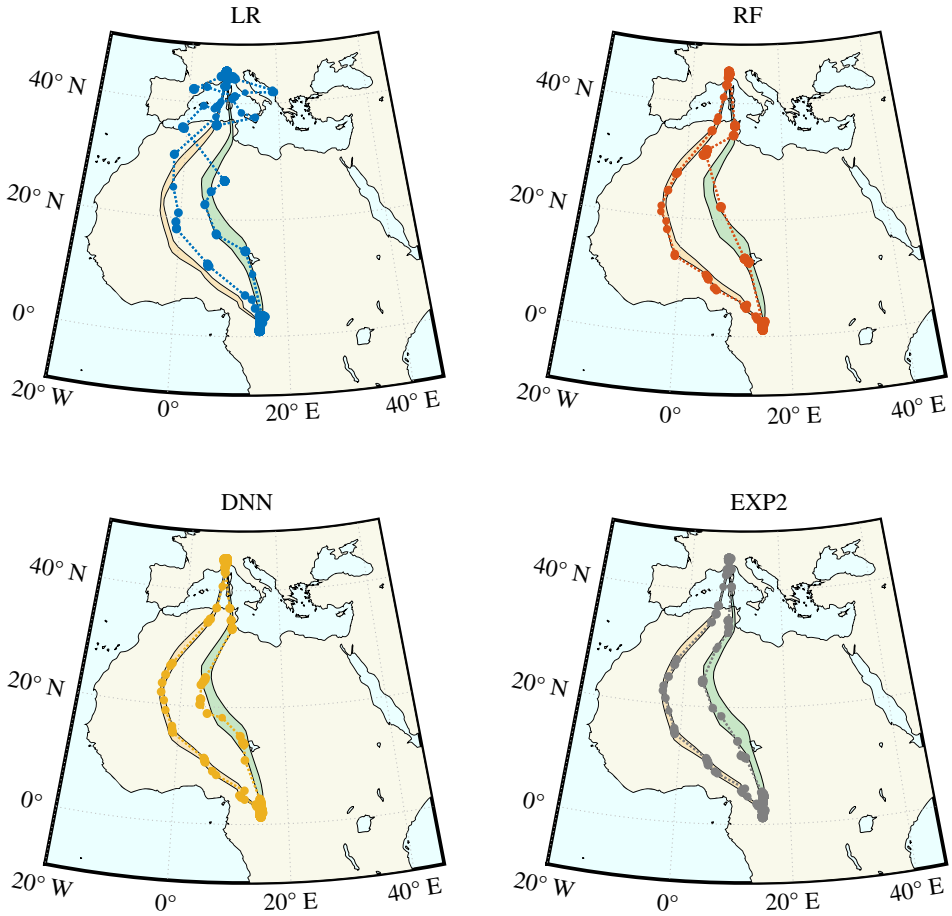

Figure 6. Migration paths estimated for one of the four target individuals using FLightR software with twilight events classified by different methods: LR (blue), RF (orange), DNN (yellow) and second classification by the expert (EXP2, grey). Filled polygons refer to first expert classification (eastern: autumn migration, western: spring migration; boundaries are obtained as mean+standard deviation of 10 repetitions of FLightR on EXP1 classification). (Online version in colour.)

Table 1. Intraclass correlation coefficient (ICC) values computed for all classification types (ALL versus ALL) and separately for each model against the first classification by the expert (EXP1). The second column shows ICC values at the level of the twilight event identifier (nested within geolocator); the third column indicates ICC values at the level of the geolocator identifier.

comparison	$ICC_{twilight\text{-}ID\ (nested)}$	$ICC_{geolocator\text{-}ID}$
ALL versus ALL	0.824	0.011
EXP1 versus LR	0.542	0.011
EXP1 versus RF	0.826	0.009
EXP1 versus DNN	0.840	0.011
EXP1 versus EXP2	0.869	0.007

or more exposed environments), which may affect the light curves on which the classification process is based.

3.2. Track comparison

Beyond comparing the classifications of twilights performed by our different ML algorithms, we aimed to assess whether ML can provide a reliable pre-filtering of twilight data for reconstructing migration routes of animals. This is crucial for assessing the utility of ML algorithms, as it would enable the most time-consuming step in geolocator data analysis to be automated. Figure 6 shows a representative route of one of the example individuals generated by FLightR (displayed as average latitude and average longitude of all particles at each twilight). The route is estimated starting from the classification operated by the ML algorithms and by the expert. The path obtained from twilights classified by the LR clearly presents estimation errors, which, together with the lower classification performance with respect to other methods exposed above, suggests that this method may be not suitable for safely automating light curve editing. By contrast, the routes generated from the classification of the RF and the DNN are very similar to those obtained from the expert classifications. We obtained similar results also for the other three example individuals (see electronic supplementary material, appendix S2).

The inconsistencies observed between the paths obtained from the RF, DNN, EXP1 and EXP2 classifications could be generated both by the different selection of twilight events and by the intrinsically stochastic nature of the Monte Carlo algorithm (i.e. the particle filter) of FLightR that can produce slightly different outputs, particularly around equinoxes, when the information provided by the light curves cannot be used to estimate latitude reliably (for further details, see [41]). In addition, it is necessary to highlight that the estimate of the positions provided by the analysis of geolocator data is affected by a significative uncertainty: on average 50 km in longitude and 200 km in latitude [42].

Table 2 shows the values of the OWD calculated by comparing the routes obtained from the classification made by LR, RF, DNN and the second classification by the

Table 2. Comparison of the OWD metric computed on the four test individuals between the routes generated by first classification by the expert (EXP1) and the other classification methods: LR, RF, DNN and our second classification by the expert (EXP2). The length of estimated paths and the computation time of each method for the four test individuals are also reported.

	OWD		path length (10^3 km)		computation time (s)	
classification	mean	s.d.	mean	s.d.	mean	s.d.
LR	4.53	1.97	15.3	2.89	<0.001	<0.001
RF	3.68	1.32	11.4	1.22	2.041	0.348
DNN	2.85	1.14	11.1	0.96	15.804	4.872
EXP2	3.40	1.60	11.1	1.10	>1200	>60

expert (EXP2) with the paths estimated through the first classification by the expert (EXP1).

A mixed-effect ANOVA of OWD with geolocator-ID as a random effect revealed significant differences according to classification types ($F_{3,9} = 6.04$, $p = 0.015$). Post hoc tests indicated that OWD of LR was significantly larger than that of EXP2 and DNN ($t_9 \geq 2.79$, $p \leq 0.021$) and marginally not significantly larger than that of RF ($t_9 = 2.10$, $p = 0.065$), while those of RF, DNN and EXP2 did not differ significantly from one another ($|t_9| \leq 1.96$, $p \geq 0.082$). Table 2 shows also the average and the standard deviation of computation time for the classification of the four test individuals (computations performed on an Intel® Core™ i3-2310M CPU, 2.10 GHz). The RF was seven times faster than the DNN, but they were both much faster than that the human expert, who needed more than 20 min per geolocator to perform the twilight classification. The small difference in computation time (few seconds) between RF and DNN is negligible when compared with the average computation time required by FLightR to estimate the migratory route of one geolocator (around 2 h in our case).

4. Conclusion

The ML algorithms proposed here allow automation of a time-consuming human task, twilight selection, that is a necessary preliminary step for estimating migration routes from geolocator data. We constructed a dataset of almost 40 000 expert-classified twilight events, on the basis of which we built and calibrated three different ML algorithms, selecting predictor features that summarize the information processed by the human classifier during his choice.

The performances of complex algorithms such as RFs and DNNs can be compared with those of the human expert, both in classification scores, repeatability and in the routes estimated by the FLightR software based on the classified twilights. By contrast, a simpler technique such as LR is not able to correctly classify the twilight events, causing highly unreliable outputs in the subsequent phase of route estimation.

For the geolocator models and bird species used in this study, it is therefore possible to automate the classification of twilight events and obtain reliable results in the reconstruction of migration routes. Further twilight measurements from other geolocator models and/or from different species may be useful for a large-scale extension of this automatic procedure in the field of migratory paths reconstruction by light measurements. Although we have no data to prove it at present, we speculate that the trained models could also be applied to other species. In fact, light data retrieved by geolocators applied on bird species with different behaviours could look quite different between one another, possibly influencing the ratio between natural and shadowed twilight events recorded by the geolocator. Yet, this does not appear to negatively affect the classifiers. On the other side, the extension to other geolocator models is more critical, since they may register light values using different sampling intervals, different value ranges and even different relationships between environmental light intensity and light measurements. For this reason, some adjustments to standardize measurements from different geolocators would be required.

In this context, the DNN may have a remarkable advantage with respect to the RF. The dataset of geolocator measurements are usually composed of a limited number of classified samples, as manual twilight selection is a very time-consuming task. In principle, the training process would have to be repeated from scratch, but the small number of records would probably turn out to be insufficient to properly calibrate complex classifiers. A DNN can take advantage of what it learned from bigger datasets, such as the one considered in this paper, thanks to a learning technique known as *fine-tuning*. In this case, the parameters of the front layers, which extract general features, are kept, whereas the parameters of the final, more task-specific, layers are re-trained on the new dataset. The same process cannot be applied to an RF, which would need to be trained again from scratch on every new dataset. Thus, while the two algorithms have comparable performances, the DNN has greater flexibility in dealing with new tasks, maintaining the knowledge extracted from previous datasets.

Eventually, a preliminary analysis (i.e. retraining models on data subsets) on how the number of labelled samples available would affect the classifiers precision has been performed. The two advanced ML algorithms (RF and DNN) could maintain similar performances to the one presented in this study, obtained with a dataset of almost 40 000 events, using just 10 000 classified twilights, which correspond to almost 20 complete geolocator tracks. However, this result is related to this specific case: algorithms trained with other geolocator devices and/or other experts classifications could in principle require a different number of twilight events.

Data accessibility. This article has no additional data.

Authors' contributions. M.P. and M.S. conceived the study under the supervision of R.C. M.P. edited the whole set of geolocators with D.W.W. and served as EXP1 and EXP2. M.S. coded and trained the ML algorithms, selecting the features together with M.P. and R.C. R.A., R.C. and N.S. coordinated the design of statistical tests and uncertainty estimation, implemented by M.P. A first draft of the manuscript was written by M.P., M.S. and R.C. All authors contributed to critically discussing results and to finalize the paper.

Competing interests. We declare we have no competing interests.

Funding. M.P. and R.C. acknowledge funding from the H2020 project 'ECOPOTENTIAL: Improving future ecosystem benefits through Earth observations' (project ID 641762).

Acknowledgements. The authors are grateful to the Swiss Ornithological Institute, in particular to Dr Felix Liechti, for providing part of the geolocator data used in this study. We also warmly thank Prof. Diego Rubolini, Dr Chiara Scandolara and several field assistants for their invaluable contribution to fieldwork during geolocator deployment and recovery. By the time this article went to the last round of review, our co-author, friend and guide, Prof. Nicola Saino, suddenly passed away—just the day after having approved our submission. His dedication to the study of behavioural ecology of birds (with special focus on barn swallows), by using all conceivable means, was not only evident to whoever had the fortune of working with him, but it was so scientifically challenging as to become irresistibly involving.

References

1. IPCC. 2014 Climate change 2014: synthesis report. In *Contribution of working groups I, II and III to the fifth assessment report of the intergovernmental panel on climate change* (eds core writing team, RK Pachauri, LA Meyer), 151 p. Geneva, Switzerland: IPCC.
2. Parmesan C. 2006 Ecological and evolutionary responses to recent climate change. *Annu. Rev. Ecol. Evol. Syst.* **37**, 637–669. (doi:10.1146/annurev.ecolsys.37.091305.110100)
3. Wormworth J, Sekercioglu CH. *Winged sentinels: birds and climate change*. Cambridge, UK: Cambridge University Press.
4. Winkler DW *et al.* 2014 Cues, strategies, and outcomes: how migrating vertebrates track environmental change. *Mov. Ecol.* **2**, 10. (doi:10.1186/2051-3933-2-10)
5. Cooke SJ, Midwood JD, Thiem JD, Klimley P, Lucas MC, Thorstad EB, Eiler J, Holbrook C, Ebner BC. 2013 Tracking animals in freshwater with electronic tags: past, present and future. *Anim. Biotelem.* **1**, 5. (doi:10.1186/2050-3385-1-5)
6. Robinson WD, Bowlin MS, Bisson I, Shamoun-Baranes J, Thorup K, Diehl RH, Kunz TH, Mabey S, Winkler DW. 2010 Integrating concepts and technologies to advance the study of bird migration. *Front. Ecol. Environ.* **8**, 354–361. (doi:10.1890/080179)
7. O'Mara OT, Wikelski M, Dechmann DKN. 2014 50 years of bat tracking: device attachment and future directions. *Methods Ecol. Evol.* **5**, 311–319. (doi:10.1111/2041-210X.12172)
8. Tomkiewicz SM, Fuller MR, Kie JG, Bates KK. 2010 Global positioning system and associated technologies in animal behaviour and ecological research. *Phil. Trans. R. Soc. B* **365**, 20100090. (doi:10.1098/rstb.2010.0090)
9. Wikelski M, Kays RW, Kasdin NJ, Thorup K, Smith JA, Swenson GW. 2007 Going wild: what a global small-animal tracking system could do for experimental biologists. *J. Exp. Biol.* **210**, 181–186. (doi:10.1242/jeb.02629)
10. Cooke SJ. 2008 Biotelemetry and biologging in endangered species research and animal conservation: relevance to regional, national, and IUCN Red List threat assessments. *Endanger. Species Res.* **4**, 165–185. (doi:10.3354/esr00063)
11. Hays GC, Mortimer JA, Ierodiaconou D, Esteban N. 2014 Use of long-distance migration patterns of an endangered species to inform conservation planning for the world's largest marine protected area. *Conserv. Biol.* **28**, 1636–1644. (doi:10.1111/cobi.12325)
12. Gilbert M *et al.* 2010 Flying over an infected landscape: distribution of highly pathogenic avian influenza H5N1 risk in South Asia and satellite tracking of wild waterfowl. *Ecohealth* **7**, 448–458. (doi:10.1007/s10393-010-0672-8)
13. Newton I. 2007 *The migration ecology of birds*. Amsterdam, The Netherlands: Elsevier.
14. Weiser EL *et al.* 2015 Effects of geolocators on hatching success, return rates, breeding movements, and change in body mass in 16 species of Arctic-breeding shorebirds. *Mov. Ecol.* **4**, 12. (doi:10.1186/s40462-016-0077-6)
15. Scandolara C *et al.* 2014 Impact of miniaturized geolocators on barn swallow *Hirundo rustica* fitness traits. *J. Avian Biol.* **45**, 417–423. (doi:10.1111/jav.00412)
16. Bridge ES, Thorup K, Bowlin MS, Chilson PB, Diehl RH, Fléron RW, Hartl P, Kays R, Kelly JF. 2011 Technology on the move: recent and forthcoming innovations for tracking migratory birds. *Bioscience* **61**, 689–698. (doi:10.1525/bio.2011.61.9.7)
17. Bridge ES, Kelly JF, Contina A, Gabrielson RM, MacCurdy RB, Winkler DW. 2013 Advances in tracking small migratory birds: a technical review of light-level geolocation. *J. Field Ornithol.* **84**, 121–137. (doi:10.1111/jofo.12011)
18. McKinnon EA, Love OP. 2018 Ten years tracking the migrations of small landbirds: lessons learned in the golden age of bio-logging. *Auk* **135**, 834–857. (doi:10.1642/AUK-17-202.1)
19. Lisovski S, Hahn S. 2012 GeoLight—processing and analysing light-based geolocator data in R. *Methods Ecol. Evol.* **3**, 1055–1059. (doi:10.1111/j.2041-210X.2012.00248.x)
20. Sumner MD, Wotherspoon SJ, Hindell MA. 2009 Bayesian estimation of animal movement from archival and satellite tags. *PLoS ONE* **4**, e7324. (doi:10.1371/journal.pone.0007324)
21. Rakhimberdiev E, Saveliev A, Piersma T, Karagicheva J. 2017 FLightR: an R package for reconstructing animal paths from solar geolocation loggers. *Methods Ecol. Evol.* **8**, 1482–1487. (doi:10.1111/2041-210X.12765)
22. Ekstrom P. 2007 Error measures for template-fit geolocation based on light. *Deep Res. Part II Top. Stud. Oceanogr.* **54**, 392–403. (doi:10.1016/j.dsr2.2006.12.002)
23. Wotherspoon S, Sumner M, Lisovski S. 2016 TwGeos: basic data processing for light-level geolocation archival tags—version 0.0-1. (rdrr.io/github/slisovski/TwGeos/).
24. Liechti F *et al.* 2015 Timing of migration and residence areas during the non-breeding period of barn swallows *Hirundo rustica* in relation to sex and population. *J. Avian Biol.* **46**, 254–265. (doi:10.1111/jav.00485)
25. Shrout PE, Fleiss JL. 1979 Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* **86**, 420–428. (doi: 10.1037/0033-2909.86.2.420)
26. Sokal RR, Rohlf FJ. 1995 *Biometry: the principles and practice of statistics in biological research*, 3rd edn, pp. 813–819. New York, NY: W.H. Freeman and Company.
27. Bates D, Mächler M, Bolker B, Walker S. 2015 Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48. (doi: 10.18637/jss.v067.i01)
28. Cox DR. 1958 The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B* **20**, 215–242.
29. Ho TK. 1995 Random decision forests. In *Proc. 3rd Int. Conf. Document Analysis and Recognition, ICDAR 1995, 14–15 August, Montreal, Canada,* pp. 278–282. Piscataway, NJ: IEEE.
30. Bengio Y. 2009 Learning deep architectures for AI. *Found. Trends Mach. Learn.* **2**, 1–127. (doi:10.1561/2200000006)
31. Biau G, Scornet E. 2015 A random forest guided tour. *ArXiv. Test* **25**, 197–227. (doi:10.1007/s11749-016-0481-7)
32. James G, Witten D, Trevor H, Tibshirani R. 2013 Tree-based methods. In *An introduction to statistical learning*, pp. 303–330. Berlin, Germany: Springer.
33. LeCun Y, Bengio Y. 1995 Convolutional networks for images, speech, and time series. In *Handbook of brain theory neural networks* (ed. MA Arbib). Cambridge, MA: MIT Press.
34. Goodfellow I, Bengio Y, Courville A. 2016 Convolutional networks. In *Deep learning* (ed. T Dietterich), pp. 321–359. Cambridge, MA; London, UK: MIT Press.
35. Hastie T, Tibshirani R, Friedman J. 2009 The elements of statistical learning. In *Bayesian*

forecasting and dynamic models. New York, NY: Springer. (doi:10.1007/978-0-387-84858-7)

36. Pedregosa F *et al.* 2012 Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830. (doi:10.1007/s13398-014-0173-7.2)
37. Chollet F. 2015 Keras Documentation. See https://keras.io/.
38. Rakhimberdiev E, Winkler DW, Bridge E, Seavy NE, Sheldon D, Piersma T *et al.* 2015 A hidden Markov model for reconstructing animal paths from solar geolocation loggers using templates for light intensity. *Mov. Ecol.* **3**, 25. (doi:10.1186/s40462-015-0062-5)
39. Kitagawa G. 1996 Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Source J. Comput. Graph Stat.* **5**, 1–25.
40. Lin B, Su J. 2005 Shapes based trajectory queries for moving objects. In *GIS '05: Proc. of the 13th ACM Int. Workshop on Geographic Information Systems, 4–5 November, Bremen, Germany*. New York, NY: ACM Press.
41. Lisovski S, Hewson CM, Klaassen RHG, Korner-Nievergelt F, Kristensen MW, Hahn S. 2012 Geolocation by light: accuracy and precision affected by environmental factors. *Methods Ecol. Evol.* **2**, 603–612. (doi:10.1111/j.2041-210X.2012.00185.x)
42. Phillips RA, Silk JRD, Croxall JP, Afanasyev V, Briggs DR. 2004 Accuracy of geolocation estimates for flying seabirds. *Mar. Ecol. Prog. Ser.* **266**, 265–272. (doi:10.3354/meps266265)